DOCUMENT RESUME

ED 281 890 TM 870 312

AUTHOR DeMauro, Gerald E.
TITLE Towards a Normative Evaluation of English Skills: An Analysis of the Appropriateness of Monolingual and Program Norms.
PUB DATE Apr 87
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bilingual Education Programs; Criterion Referenced Tests; Elementary Secondary Education; *English (Second Language); *Evaluation Criteria; *Language Skills; Limited English Speaking; Listening Comprehension Tests; *Norm Referenced Tests; Pretests Posttests; Program Evaluation; Reading Tests; *Test Norms; Writing Evaluation
IDENTIFIERS *Language Assessment Battery; Monolingual Students; New York (New York); Normal Curve Equivalent Scores; Percentile Ranking

ABSTRACT

Evaluation of bilingual education programs is limited by the absence of appropriate standards against which program success may be gauged. Normative and criterion models have been used, inadequately. The current paper revisits the normative model, and examines the relative advantages of using norms based on the performances of monolingual-curriculum students or program participants. Data from New York City's standardization administration of the Language Assessment Battery (LAB) were analyzed. The results show that use of monolingual-curriculum norms advantage program students in listening skills, and also advantage students in higher grades. Elsewhere, program participant norms afford greater advantage to the program students. These findings are consonant with language acquisition theory (Cummins, 1979). (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281890

TOWARDS A NORMATIVE EVALUATION OF ENGLISH SKILLS: AN ANALYSIS OF THE APPROPRIATENESS OF MONOLINGUAL AND PROGRAM NORMS

Gerald E. DeMauro
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. E. DeMauro

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 870 312

Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 1987

BEST COPY AVAILABLE

ABSTRACT

Evaluation of bilingual education programs is limited by the absence of appropriate standards against which program success may be gauged. Normative and criterion models have been used, inadequately. The current paper revisits the normative model, and examines the relative advantages of using norms based on the performances of monolingual-curriculum students or program participants. Data from New York City's standardization administration of the LAB were analyzed. The results show that use of monolingual-curriculum norms advantage program students in listening skills, and also advantage students in higher grades. Elsewhere, program participant norms afford greater advantage to the program students. These findings are consonant with language acquisition theory (Cummins, 1979).

Towards a Normative Evaluation of English Skills: An Analysis of the Appropriateness of Monolingual and Program Norms

Gerald E. DeMauro
Educational Testing Service

Evaluation of bilingual education programs is complicated by an inability to project how students would perform without the program (DeMauro, 1983). Two solutions to the dilemma with wide currency follow.

Normative Models

Normative models are based on the Chapter I reporting system (Tallmadge and Wood, 1976). Many state education agencies employ them (New York and New Jersey, for example), and it requires that pretest performance be referenced to test norms. From the pretest, a posttest score is projected that maintains each student's position relative to the test norms. In essence, the normative sample becomes a proxy control group.

For evaluation of skills in content areas (reading, math, etc.), the model assumes that changes in English proficiency are not related to measurement of content achievement (DeMauro, 1983). That is, if a student's norm-referenced score changes between fall and spring, the change is assumed to reflect changes in reading or math skills, rather than changes in English abilities. Moreover, the model is inadequate for students undergoing transition in the language

of instruction (DeMauro, in press) unless provisions are made to estimate the effect of transition, itself, on academic growth.

When English acquisition is evaluated, the model employs either of two types of norms: those based on the performance of limited English proficient (LEP) students in bilingual education or English as a second language (ESL) programs or those based on the performance of students in a mainstream monolingual-curriculum.[1] The first type fails to provide a no program expectation, since program students are compared to program students. The second type assumes that program students and monolingual-curriculum students with the same pretest score relative to the norms would grow the same amount without the program. The interdependence of first and second language skills (Cummins, 1978) is thus assumed to have no effect.

Because the model fails to separate from error variance both the level of first language skills and the interaction of first language skills with pretest to posttest growth in

---

[1]These norms are sometimes called "English Proficient" norms, referring to the skills of the normative group. They are also sometimes called "monolingual", which refers to the curriculum of the normative group.

English, it is a conservative model. Presumably, the contributions of these real effects to error variance reduce the power of the model. Moreover, since it is precisely English skills that differentiates normative and program students, the overlap in score distributions is likely to be small, severely limiting the discriminating power of the norms.

Criterion-Referenced Models

Criterion-referenced models examine the skills to be measured and determine what level of performance would be necessary for program success. Measurements are then made to determine whether students have achieved this level.

Analysis of program exit rates is one application of such a model. States or programs may decide that 75 percent of L.E.P. students should meet all exit requirements by the fifth year of program participation. Actual exit rates are then compared to projected rates.

The obvious difficulty with this model is, again, the no-program expectation. It is not possible to project what percentage of program students would have attained the criteria without the program.

One compromise solution is to establish a criterion of mastery of certain skills on the posttest based on a projection of what skills should be mastered in order to satisfactorily exit the program after a given number of years. In practice, one would look at the skills mastered on the pretest, and from that, develop a series of intermediary goals and a final goal that accounts for the abilities expected of students in the mainstream monolingual classroom at the appropriate grade level. A latent trait model may be employed to make these projections (DeMauro, 1985). That is, the probability of mastering certain skills can be related through item response theory, to score levels of tests. Projected score levels may then be compared to actual score levels.

Neither the normative nor criterion models is sufficient for evaluating bilingual education, since neither provides the required no-program expectation. Any evaluation of bilingual education must account for progress towards the program's goals and relate that progress to elements of the program. For example, in many states, transitional bilingual education has the goal of enabling limited English proficient students to participate in the school district's monolingual curriculum. Both criterion and normative models assess the students' readiness for such mainstreaming, but neither, by itself, can relate that readiness to the program. The

criterion model can make test scores interpretable in terms of observed deficits of program students. The normative models make test scores interpretable either in terms of how they compare to scores obtained in similar programs (program participant norms) or in terms of how they compare to skills distributions of monolingual-curriculum students.

The current paper examines the ability of normative models in terms of the advantages of using monolingual-curriculum or program participant norms to evaluate English skills acquisition. In particular, an examination is made of whether students of given pretest skills levels or grade levels are disadvantaged by the use of monolingual-curriculum norms. The analyses also examines whether any advantages of these norms are greater in some English skills areas (e.g., listening) than others (e.g., writing).

Methods

Procedures. As part of the 1982 standardization study of New York City's Language Assessment Battery, the instrument was administered to two groups of students, "target" students and students in the city's monolingual-curriculum (Abbott, 1985). The "target" students were participants in New York's bilingual and English as a Second Language programs (ESL).

Subtest and total test program ("target") norms were made available by New York City's Office of Assessment. Monolingual-curriculum (the Office of Assessment calls these "English Proficient") norms were similarly available. Each set of norms was available in all grade levels.

Instrument. The Language Assessment Battery (LAB), has four levels, encompassing grades kindergarten through second, third through fifth, sixth through eighth, and ninth through twelfth. Spanish and English versions of the battery are available. Level 1, has a Listening and Speaking and a Reading and Writing subtest. The remaining three levels have Speaking, Listening, Reading, and Writing subtests. Because the Speaking subtest is individually administered, it is typically not used. Consequently, only the Listening, Reading, and Writing subtests were analyzed. Moreover, since Level 1 has unique subtests, only Levels 2-4 were analyzed. Finally, since the focus of the study is differences between the utility of monolingual-curriculum and program-participant norms, the analysis is restricted to the English version of the Battery.

Data. The raw score corresponding to every tenth percentile (tenth, twentieth, ..., ninetieth) on the program participant norms was determined in each grade level

from third through twelfth, and for each of the subtests of listening, reading, and writing. This procedure was repeated for the fall and spring distributions. The spring raw score corresponding to each percentile represents the score each student who achieves that same percentile in the fall must achieve to maintain his/her position relative to the norming group. Students who score higher than the spring score have outgained students who had comparable pretest scores. This analysis is based on the Chapter 1 normative model (Tallmadge and Wood, 1976).

Once these raw scores were determined, they were converted to monolingual-curriculum percentiles through reference to the monolingual-curriculum norms tables. Differences from fall to spring in this metric represent gains (or losses) made by program students who did not change with respect to the program-participant norms. When the spring score is higher than the fall score, program students who comprised the "target" norms gained more than the students who comprised the monolingual-curriculum norms. On the other hand, when the fall score is higher, the "target" norms students gained less than the monolingual-curriculum sample.

Finally, the monolingual-curriculum percentiles were converted to Normal Curve Equivalents (NCEs), an interval level of measurement. This enabled computation of the amount

of gain at each grade level, each subtest, and each of the nine ability levels and comparing differences in gains across these variables. This is not possible using percentiles, an ordinal level of measurement, since differences between any two percentiles are not comparable to differences between any other two percentiles.

Analysis

The amount of gain (or loss) in NCEs for scores converted to monolingual-curriculum NCEs was smoothed at each of nine percentiles, obtaining the weighted average of the gain scores (weighted by the NCEs corresponding to the percentile) with those of the adjoining above and below weighted by the score.[2] The smoothing was done to control for score variations related to sampling distortions. At the 10th percentile class only the scores from the 20th percentiles was used to smooth the gains, but at the 90th percentile, scores from both the 80th and 99th percentiles were used to smoothe the gains. This method was adapted from one described by Livingston and Zieky, (1982).

---

[2]The weighting formula was: $$\frac{NCE_1\ (G_1) + NCE_2\ (G_2) + NCE_3\ (G_3)}{NCE_1 + NCE_2 + NCE_3}$$

The analysis examined the NCE gains within each grade and subtest as a function of the program participant (target) norms percentiles. Particular attention was given to the direction of gain to determine places in the target distribution where use of monolingual-curriculum norms would advantage the program, that is where the growth of program students is greater than the growth of monolingual-curriculum students.

Results

Similarities. For each test level in each subtest, the smallest differences are found in lower ability students (lower percentiles) in Level 4 of the Battery than in Levels 2 or 3 of the Battery. That is, in upper grades, ability program participants were more similar to monolingual-curriculum students in English growth, while in lower grades, higher ability program participants were more similar. For example, for Listening, the smallest difference was in the fiftieth percentile in Level 2, but in the 30th in Levels 3 and 4. For Reading, the smallest difference in Level 2 was in the 80th percentile, in the 80th - 90th in Level 3, and in the 50th and 80th in Level 4. In Writing, the smallest difference is in the 90th percentile for Levels 2 and 3, but in the 50th in Level 4. Simarily, for the total test score, the smallest difference is in the 90th percentile for Level 2, the 80th for Level 3, and the 40th for Level 4.

Table 1 shows that for Listening, the smallest mean differences in each test level were achieved in fourth, sixth, and eleventh grades. For Reading the smallest mean differences were achieved in third, sixth, and eleventh grade. For Writing, the smallest mean differences were achieved in fifth, eighth, and ninth grades, and for the total test score in third, seventh, and eleventh grades. Finally, in the lower grades, the smallest differences tend to be in Listening, but the smallest differences tend to be in Reading in higher grades.

Differences - Level 2. In Listening skills, program participants from the tenth to the fiftieth target norms percentiles gained less (negative mean gains) than comparably-skilled monolingual-curriculum students. The opposite was true of higher ability program participants, who gained more than comparably-skilled monolingual-curriculum students.

Program participants from the tenth to eightieth target percentiles also gained less than comparably-skilled monolingual-curriculum students in Reading, and all program participants gained less than comparably-skilled monolingual-curriculum students in Writing and on the total test. In all cases, the mean fall to spring gains were larger for higher ability (higher target percentiles) program students.

Differences - Level 3. Program participants from the tenth to twentieth and in the eightieth target percentiles made smaller gains than comparably-skilled monolingual-curriculum students in Listening skills. In other levels of ability, the program participants outgained monolingual-curriculum students.

Program participants from the tenth to the eightieth percentiles gained less than monolingual-curriculum students in Reading and on the total test. Program participants from the tenth to seventieth percentiles also gained less in Writing. In Reading and Writing and on the total test, mean fall to spring gains were larger for higher ability program students.

Differences - Level 4. In Listening, program participants gained more than monolingual-curriculum students in all but the tenth to the twentieth percentiles. In Reading, program participants gained less than monolingual-curriculum students in the tenth and sixtieth to seventieth target percentiles. Similarly, in Writing, program participants made smaller gains in the tenth and fiftieth to seventieth percentiles. Finally, program participants in the fortieth to ninetieth percentiles gained less than comparably-skilled monolingual-curriculum students on the total test score.

Overall by Grade. Table 1 shows that program participants made the smallest gains with respect to monolingual-curriculum students in the highest grade within each test level. The only exception to this, is in Level 4, Reading, where the smallest gain was made by eleventh graders.

Discussion

It appears that in the lower grades the smallest differencies were in Listening, and in the higher grades, the smallest differences were in Reading. If one considers that the mode of instruction becomes increasingly reliant over grade levels on reading skills and less reliant on listening skills it is clear that the mode of instruction influences language growth or is at least atuned to it. Therefore, the amount of growth in these areas for the two groups is similar as is the curricular demand for growth.

Overall, in all test levels, use of monolingual-curriculum norms only advantages program students in the higher levels of abilities or Listening skills. Fall to spring gains were generally positive on this subtest or in these grades. This means that program students are only more likely to show gains if monolingual-curriculum norms are used to assess Listening or if they are used to assess skills of high school students. By the time students have developed only low to mid-level English skills, relative to their peers,

they are beginning to grow more quickly than monolingual-curriculum students. Perhaps this is due to the availability of social language (listening) skills already developed in another language (Cummins, 1979). As these social skills develop early, they are available to all program students by the time they reach third grade.

In Reading, Writing and the total test, in Levels 2 and 3 higher skills are required before students can avail themselves of skills developed in other languages.

In Level 4 of the test, program participants probably have academic language skills already well-developed in their native languages. Drawing on this reservoir, they are able to outdistance poorly-skilled monolingual-curriculum students without similar linguistic resources. Middle range monolingual-curriculum students do have some academic language skills. In the upper ranges, program participants again outdistance monolingual-curriculum students, because they have these academic language skills in two languages. The interaction of these two sources of skills provides a clear linguistic advantage.

In conclusion, monolingual-curriculum norms advantage program students who are tested in social language abilities. They also advantage students in higher grades, for the most part. Since these norms have great meaning in terms of the student's abilities to succeed in the monolingual-curriculum, they may prove most useful for oral assessments and for students in higher grades.

Table 1

Mean Smoothed Gains of
Students who Comprised New
York City Program Participant
Norms in Monolingual-Curriculum
NCEs by Grade Level and Subtest

		Subtest			
	Grade Level	Listening	Reading	Writing	Total
	3	1.38	-2.14	-2.79	-1.29
Level 2	4	0.44	-3.08	-4.47	-3.46
	5	-0.51	-2.47	-2.71	-3.51
	6	0.58	-0.14	-0.98	-0.86
Level 3	7	1.10	-1.23	-1.31	-0.04
	8	-1.22	-2.61	-0.33	-3.36
	9	1.41	0.18	-0.22	1.20
	10	1.17	0.32	0.72	1.06
Level 4	11	0.03	-0.04	1.53	-0.23
	12	-0.43	0.13	-1.06	-1.04

REFERENCES

Abbott, M. M. Theoretical considerations in the measurement of the English-Language Proficiency of limited-English-Proficient students. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, April 1, 1985.

Cummins, J. Linguistic interdependence and the educational development of bilingual children. Review of Educational Research, 49 (2), 1979, 225-251.

DeMauro, G. E. Models and assumptions for bilingual education evaluation. Bilingual Journal, Winter, 1983, 8-12.

DeMauro, G. E. Issues in the placement of limited-English students and in the evaluation of their instructional programs. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 1, 1985.

DeMauro, G. E. Evaluation and equivalence of measures across languages: issues of homogeneity. In S. Seidner (Ed.) Issues of Language Assessment Volume III, Springfield, IL, Illinois State Department of Education (in press).

Tallmadge, G. K. and Wood, C. T. User's Guide : ESSA Title I Evaluation and Reporting System, Mountain View, CA, RMC Corporation, 1976.